UCLA Statistics

scReadSim: a Single-cell RNA-seq and ATAC-seq Read Simulator

CMCF 2022 Annual Symposium on Multiscale Cell Fate

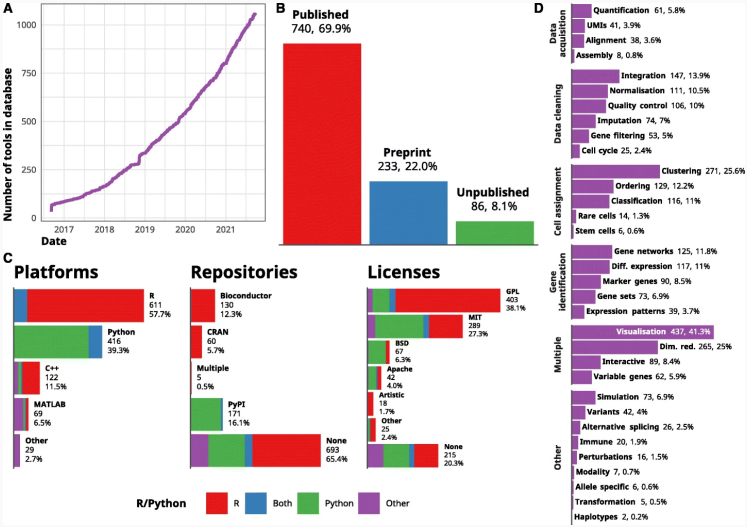
Guan'ao Yan, Jingyi Jessica Li

October 25, 2022

The Junction of Statistics and Biology
Department of Statistics
University of California, Los Angeles

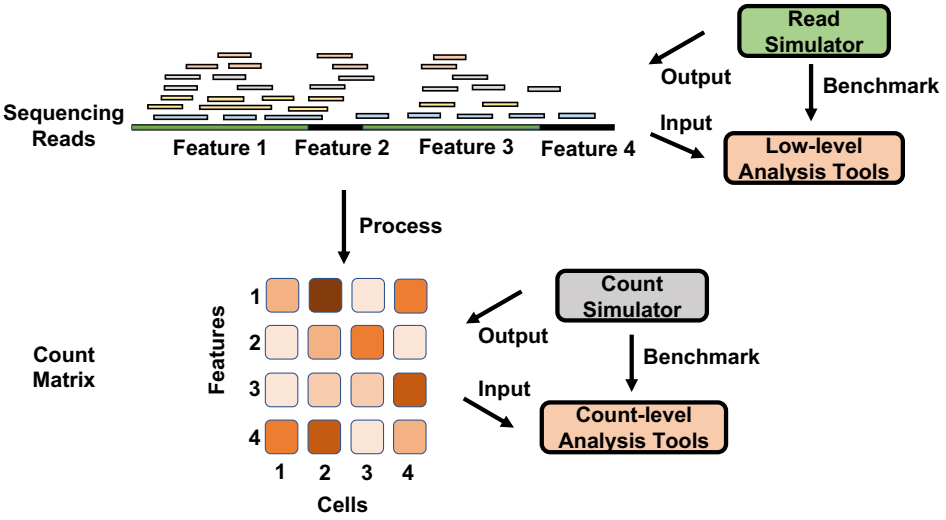
http://jsb.ucla.edu

Over 1000 tools developed for single cell RNA-seq data



Luke Zappia and Fabian J Theis. Over 1000 tools reveal trends in the single-cell rna-seq analysis landscape. Genome biology, 22(1):1–18, 2021.

Analysis tools for single-cell sequencing technologies



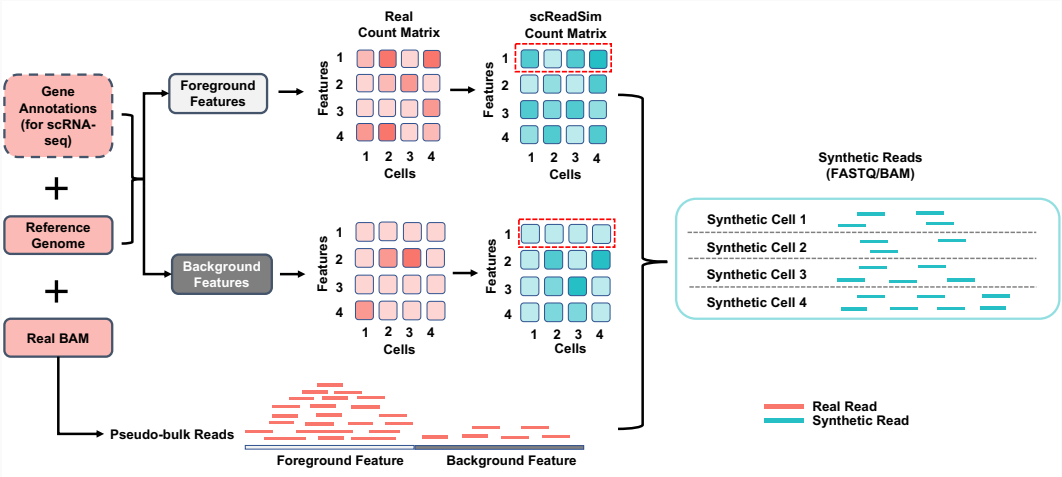
scReadSim

We here propose
scReadSim: a single-cell RNA-seq and ATAC-seq read simulator.

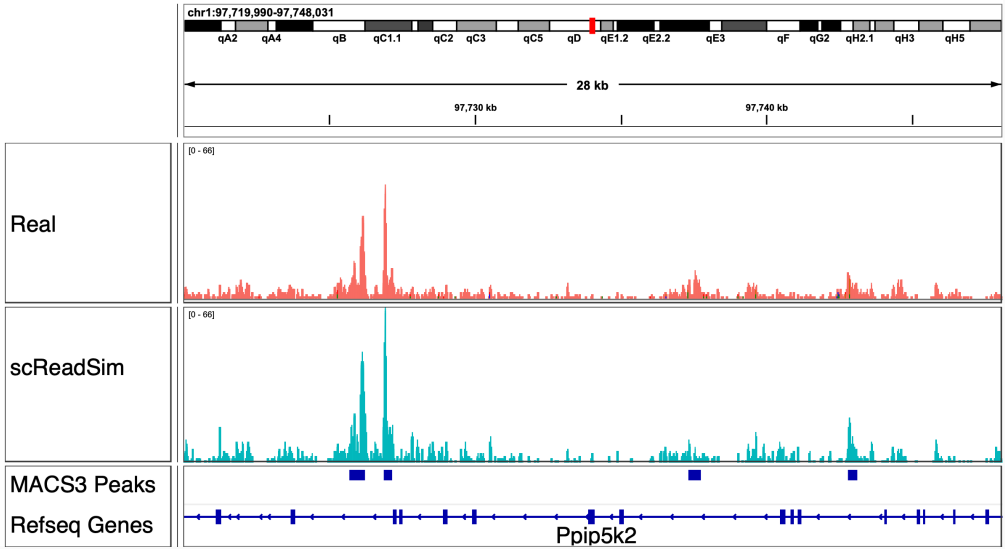
Advantages:

- aiming for **two omics data**: scRNA-seq and scATAC-seq
- simulating **realistic reads** for synthetic cells
- **flexibility**: varying cell
  number and sequence depth; user-specified peaks for scATAC-seq

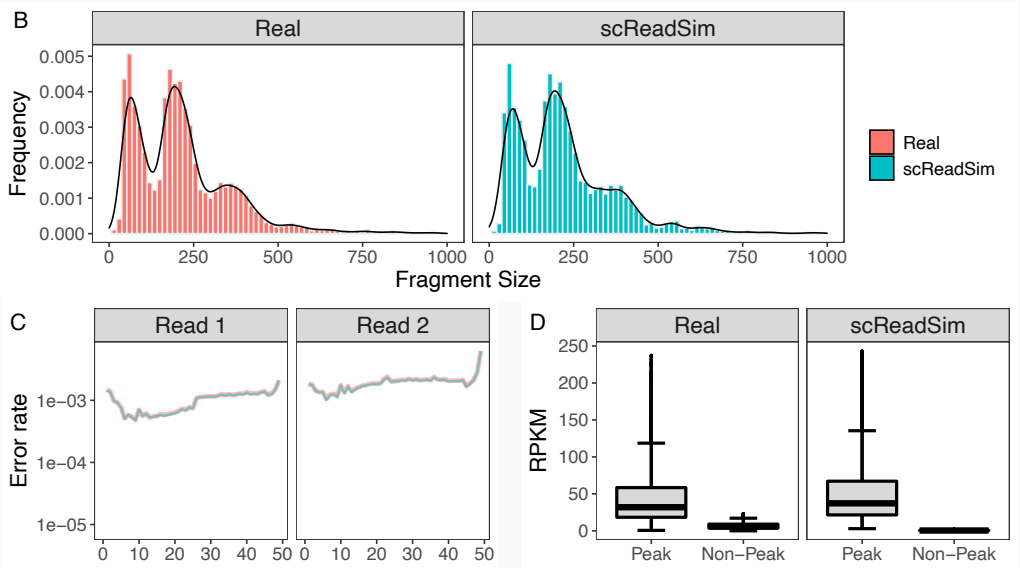
Workflow for scReadSim



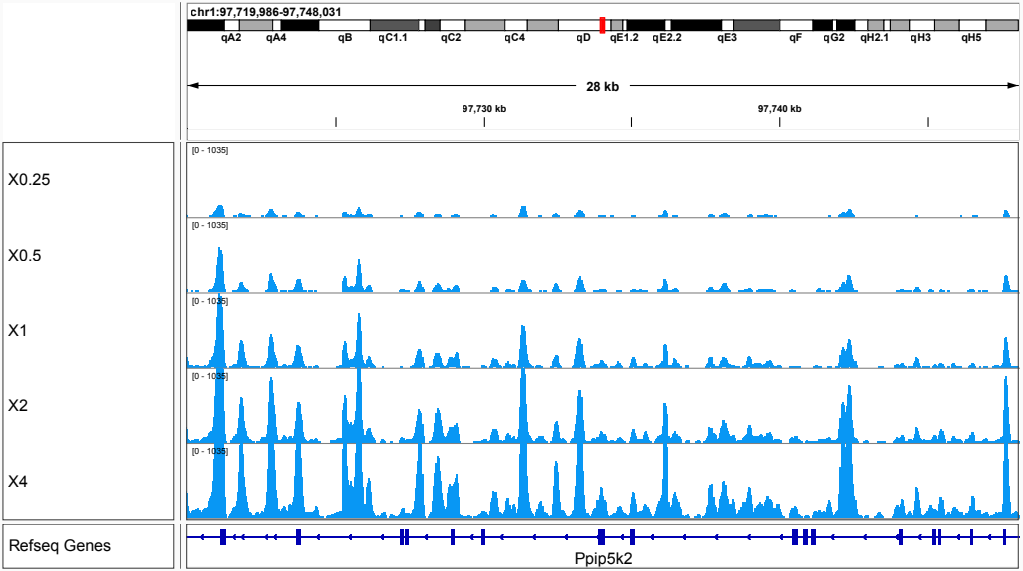
Verification: scReadSim synthetic cells mimics real cells in read level



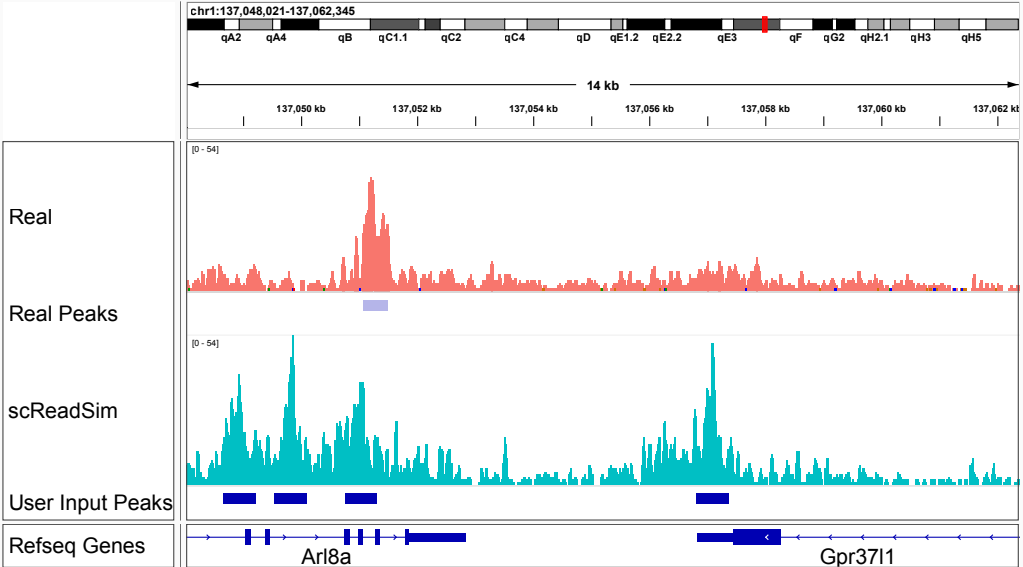
Verification: scReadSim synthetic cells mimics real cells in read level



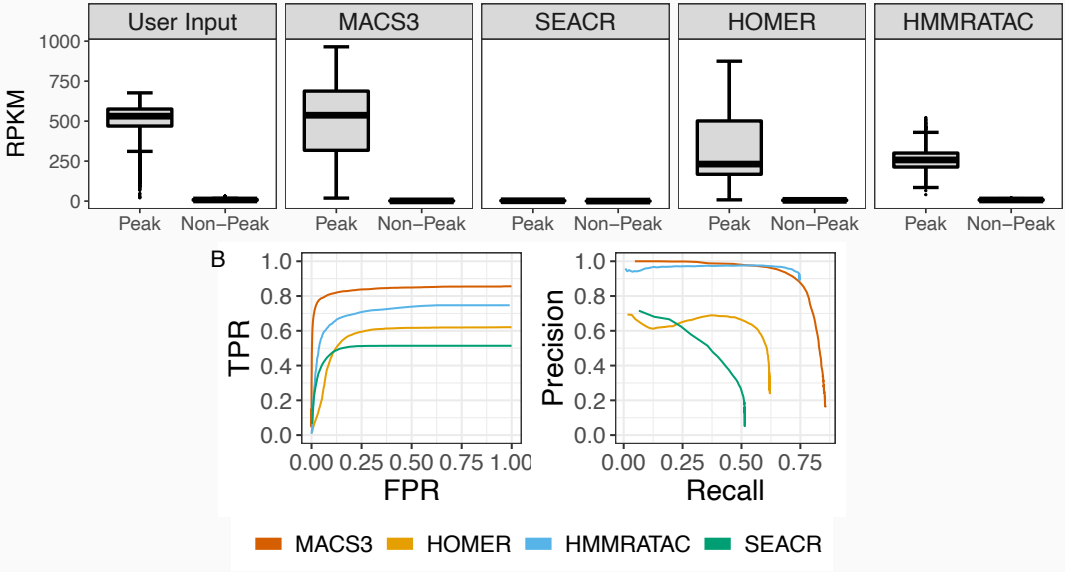
Flexibility: varying cell number and sequence depth



Flexibility: user-specified peaks for synthetic scATAC-seq data



Application: benchmark study of scATAC-seq peak calling methods



Materials

Preprint: Guanao Yan and Jingyi Jessica Li. "scReadSim: A Single-Cell Multi-Omics Read Simulator." bioRxiv. https://doi.org/10.1101/2022.05.29.493924.

Python package: http://screadsim.readthedocs.io/
Github page:
https://github.com/JSB-UCLA/scReadSim

Backup slides

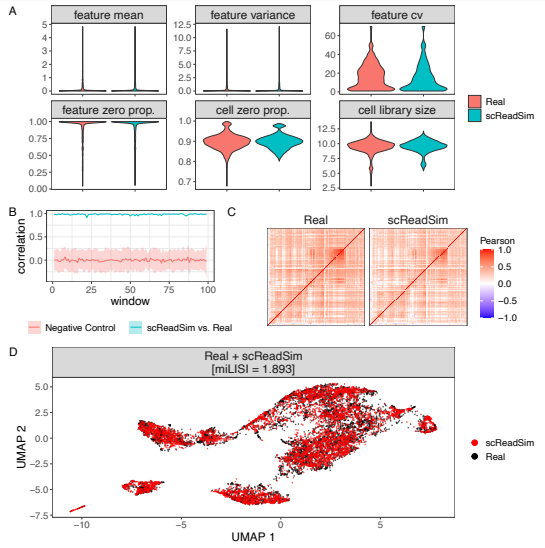
Verification 1: scReadSim synthetic cells mimics real cells in read level

- Read level
  - A. k-mer specturm
  - B. fragment length distribution
  - C. read error per base
  - D. Reads Per Kilobase Million (RPKM)
  - E. Read coverage comparison through the genome browser.
- Peak level
  - A. Venn diagram of peak calling results using MACS3 on both scReadSim synthetic and real BAM files.
  - B. True positive rate (TPR) vs. false positive rate (FPR) and false discovery rate (FDR) vs. TPR by taking real peaks as truths.
  - C. Read coverage comparison through the genome browser.

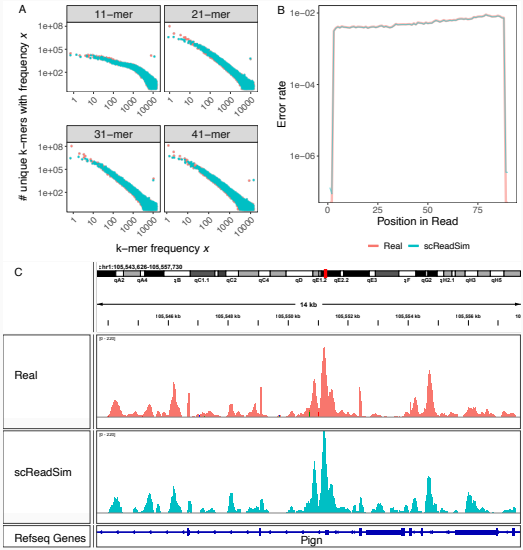
Verification 2: scReadSim synthetic cells mimics real cells in read count level

- Count level
  - A. Summary statistics: **Features (rows)** mean, variance, coefficient of variance, sparsity; **Cells (columns)** sparsity, library size.
  - B. Correlation between synthetic and real counts of windows. Each window include the same number of adjacent foreground features along the reference genome.
  - C. Correlation among top 100 expressed features.
  - D. Visualization of dimension reduction results using UMAP.

For B, we also include a control group: for each real window, we randomly select 100 real window from the remaining to calculate the correlation. The mean and standard deviation are shown as a contrast.

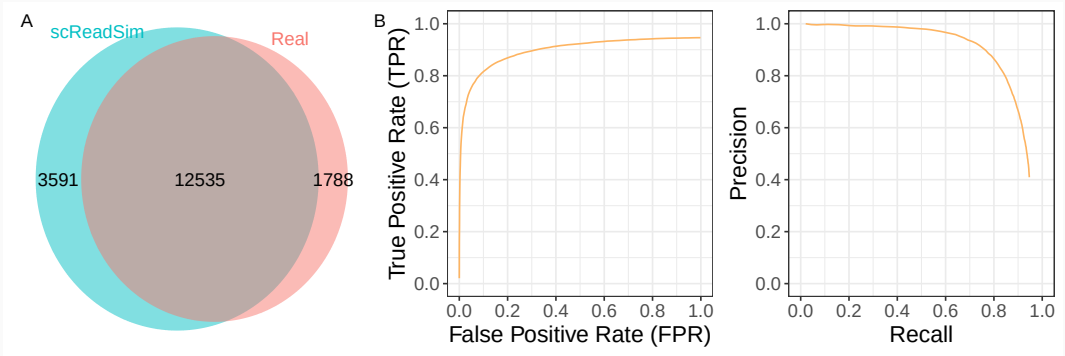


Synthetic reads generated by scReadSim resembles real 10x scRNA data in count level.

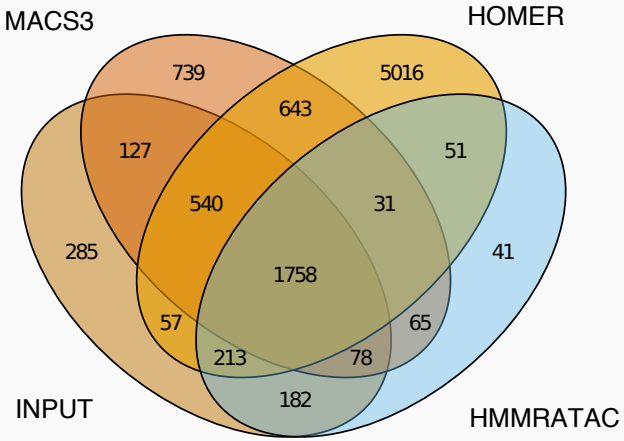


Synthetic reads generated by scReadSim resembles real 10x scRNA data in read level.

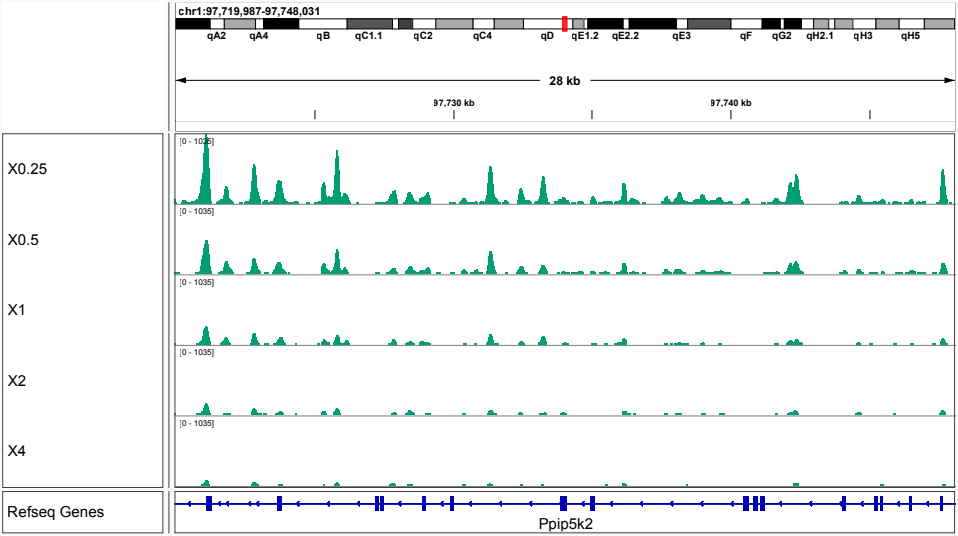
Verification: scReadSim synthetic cells mimics real cells in read level



Application 1: Benchmark study of peak calling methods



Flexibility: user-specified synthetic cell number and sequence depth



Application 2: benchmark of scRNA-seq deduplication methods